DOCUMENT RESUME

ED 356 259 TM 019 684

AUTHOR Olson, Jeffery E.
TITLE Least Principal Components Analysis (LPCA): An Alternative to Regression Analysis.
PUB DATE 92
NOTE 10p.; An earlier version of this paper was presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Error of Measurement; *Factor Analysis; Goodness of Fit; *Mathematical Models; *Maximum Likelihood Statistics; *Regression (Statistics); Research Methodology; Statistical Inference; Statistical Significance
IDENTIFIERS Bootstrap Methods; Confidence Intervals (Statistics); Eigenvalues; *Least Principal Components Analysis

ABSTRACT

Often, all of the variables in a model are latent, random, or subject to measurement error, or there is not an obvious dependent variable. When any of these conditions exist, an appropriate method for estimating the linear relationships among the variables is Least Principal Components Analysis. Least Principal Components are robust, consistent, and sufficient maximum likelihood estimates of the best total linear fit to observed data. They are more appropriate than regression estimates when the smallest eigenvalue exists and is distinct from the next smallest, and the variability to minimize is in more than one variable or when multicollinearity is a problem. They are as easy to compute as are common principal components because they are principal components. T. W. Anderson (1963) provides a theory of inferential statistics for Principal Components that can be used in computing significance levels and confidence intervals for least principal components as well. Bootstrap approaches have also been developed. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JEFFERY E. OLSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED356259

Least Principal Components Analysis (LPCA): An Alternative to Regression Analysis

Jeffery E. Olson
St. John's University, New York City

Copyright 1992

Draft--Do not copy, quote or otherwise use without the permission of the author. An earlier version of this paper was presented at the Annual Meeting of the American Educational Research Association, San Francisco, 24 April 1992.

TM019684

BEST COPY AVAILABLE

Regression analysis is a powerful statistical tool; however, there are times when its assumptions are inappropriate or too restrictive. One often inappropriate assumption is that the only latent, random, or erroneously measured variable is the dependent variable. Often all of the variables in the model are latent, random or subject to measurement error, or there is not an obvious dependent variable. When any of these conditions exist, a more appropriate method for estimating the linear relationships among the variables is Least Principal Components Analysis (LPCA).

In 1901, Carl Pearson noted that the linear model that provides the best least squares fit to observed data is the model that has for its coefficients the eigenvector associated with the smallest eigenvalue of the covariance matrix. These are the Least Principal Components coefficients. Pearson's paper provided the basis for conventional principal components analysis, but its full implications for an alternative to Regression apparently have not previously been recognized.[1]

Why Use LPCA?

LPCA is a relatively straightforward method of estimation that provides the best overall fit to the observations. It does not make a priori assumptions about error, latency or randomness that are as restrictive as Regression; and yet, the two methods are conceptually very similar.

When Is LPCA Used?

There are at least four circumstances when it would be sensible to use LPCA instead of Regression Analysis. First, sometimes it is more sensible to minimize the unexplained variability of all the observed variables rather than just a dependent variable. For example, multiple-ouput production models in economics are often defined in implicit functional form--without a dependent variable. These might sensibly be estimated through LPCA. Second, sometimes it is not wisest to assume in advance of estimation the direction of the randomness, latency or error that is to be minimized. Regression Analysis assumes that such variability lies only in the dependent variable. Third, LPCA can be used to check the price that is paid in explanatory value for using Regression Analysis. Regression estimation is a limiting case of LPCA (as I later explain) so the results from LPCA and Regression can be compared to determine how much the overall error is increased in Regression. Fourth, coefficient instability from multicollinearity in Regression Analysis can sometimes be overcome through LPCA because the covariance assumptions of the two methods differ. If the covariance assumptions of LPCA are satisfied, then LPCA provides an alternative to Regression that avoids instability from multicollinearity.

What Is LPCA?

Coefficients

LPCA, like regression analysis, estimates the coefficients of an equation that fits a linear model of observed phenomena. The elements of a vector orthogonal to a linear space equal the coefficients of the equation of the space (Figure 1) as is demonstrated in any elementary text of analytical geometry (for example, see Fuller, 1967, pp. 195-204).

---

[1]In 1989, I developed the Least Principal Components approach under the direction of William F. Massy and Michael S. Garet without being aware of the contents of Pearson's paper. I have searched widely in the multivariate literature for a paper that refers to this linear modeling approach. It makes so much intuitive sense that I assumed someone else had already explored its implications, but I have not yet found one.

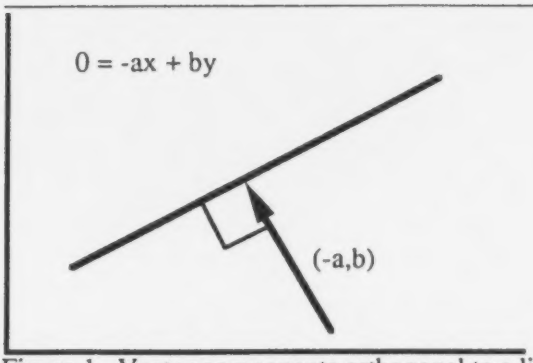

Figure 1. Vector components orthogonal to a linear surface equal the coefficients of the linear surface

The eigenvector associated with the smallest eigenvalue of the covariance matrix is the vector orthogonal to the space that minimizes the total unexplained variance of the model (for example, Flury, 1988). Figure 2 presents this in two dimensions.

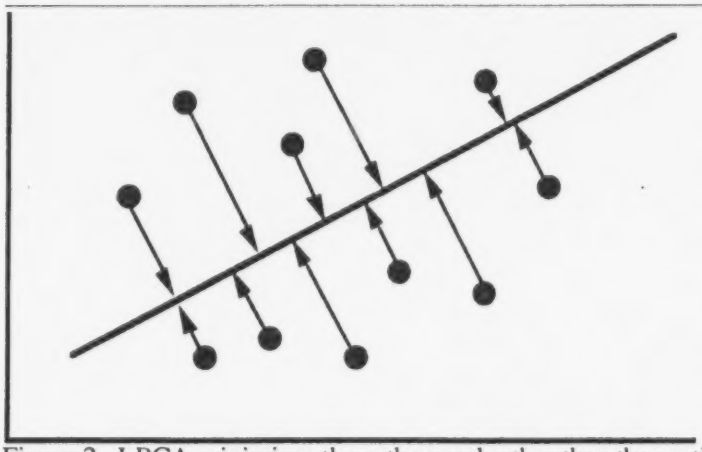

Figure 2. LPCA minimizes the orthogonal rather than the vertical distance from the observations

The elements of this eigenvector are the Least Principal Components coefficients because they are associated with the principal components of least variability--the smallest possible dimension of error.

References to Common Principal Components in the statistics literature are relevant to LPCA; Flury (1988) provides an excellent overview. LPC estimates not only provide the linear coefficients of best fit to the model (Pearson 1901); but like other principal components, they are consistent, efficient, and robust estimators when the distribution of the observations satisfy assumptions of normality and the smallest eigenvalue is distinct from the next smallest (Flury, 1988, pp. 14 and 20).

A two dimensional model provides a good view of both the similarity and difference between a linear estimate from regression analysis and one from Least Principal Components (Figure 3).

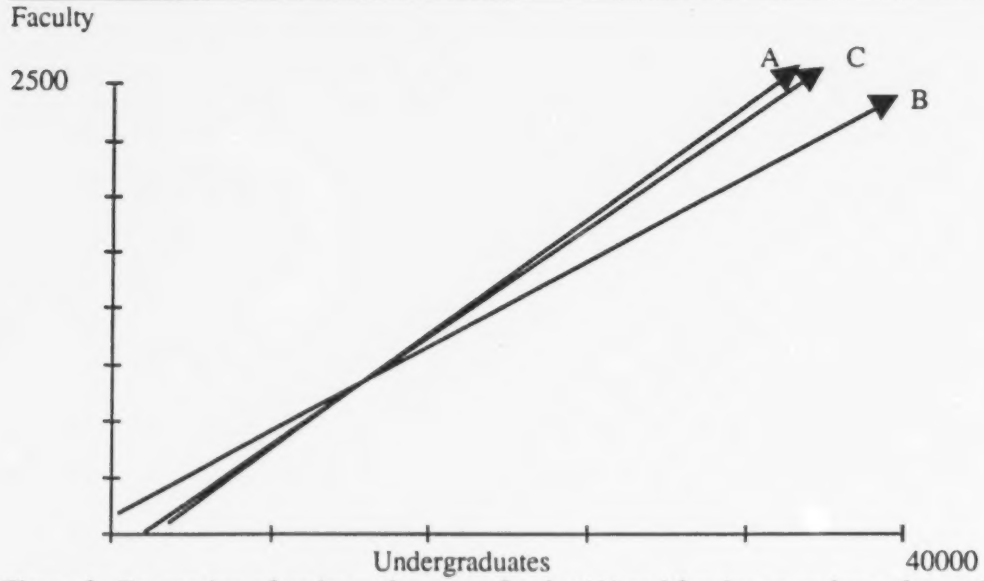

Figure 3. Regression of undergraduates on faculty (A) and faculty on undergraduates (B), and LPCA of faculty and undergraduates (C).

Figure 3. shows the number of fulltime faculty members plotted against the number of undergraduate students for the universities that awarded doctoral, first professional and bachelor's degrees in the 1982-1983 academic year. If you assume that the number of faculty members determines the number of undergraduates and that the variability of the model is only in that measurement, then you should regress the number of undergraduate students on the number of faculty members (line A), minimizing the sum of vertical squared distances from the observations to the estimation line. On the other hand, if you assume that the number of undergraduates determines the number of faculty members and that the total variability in the model is in that measurement, then you should regress the number of faculty members on the number of students (line B), minimizing the sum of horizontal squared distances from the observations to the estimation line. But, if you assume that they determine each other, neither one being the appropriate dependent variable--or variability being in more than the dependent variable and total fit making sense--then you should use LPCA (line C), minimizing the total squared distance from the observations to the estimation line.

Our three different approaches to estimation result in three different sets of estimates (Table 1), normalized to facilitate comparisons; although, rounding hides the differences between two of the sets. Regressing undergraduates on faculty results in coefficients that round to the same thousandths as the LPCA estimates. This means that the linear combination of the variables that is the dimension of least unexplained variability is very nearly the variable of undergraduates itself. On the other hand, regressing choosing a different dependent variable for regression results in a different result. The choice of dependent variable can have considerable impact on Regression estimates.

Table 1
A Comparison of Regressions with Two Different Dependent Variables and LPCA

	Regression		LPCA
	Undergraduates on Faculty A	Faculty on Undergraduates B	C
Faculty	1.000	1.000	1.000
Undergraduates	.054	.066	.054

The relative explanatory value of the two methods in terms of overall fit to the data can be compared through the proportion of the generalized variance of the all of the variables that is explained by each model. The proportion explained is one minus the proportion not explained. For both LPCA and Regression, the proportion of variance not explained is the variance of the respective error term, **e** and **e***. For LPCA, **e** equals the smallest eigenvalue:

$$1 - \frac{\mathbf{e}'\mathbf{e}}{(n)\sum_{i=1}^{p} l_i} = 1 - \frac{l_p}{\sum_{i=1}^{p} l_i},$$

where l_i is the ith eigenvalue of the covariance matrix of all variables, l_p is the smallest eigenvalue, and n is the number of observations. For Regression, the proportion explained equals

$$1 - \frac{\mathbf{e}^{*\prime}\mathbf{e}^{*}}{(n)\sum_{i=1}^{p} l_i}.$$

The proportions for LPCA and Regression equal each other when the smallest dimension of unexplained variance is in the dependent variable in the Regression.

Fitted or Latent Values

The latent values of the variables in LPCA are the projections of the observations of all of the variables onto the p-1 dimensions of the largest eigenvalues:

$$\hat{\mathbf{Z}} = \mathbf{Z}\mathbf{A}_1''\mathbf{A}_1,$$

where $\hat{\mathbf{Z}}$ is a set of fitted values corresponding to the mean-centered matrix of observations **Z**, and $\mathbf{A}_1$ is a matrix of eigenvectors of the covariance matrix with zeroes substituted for the last eigenvector. The latent values for regression analysis are the projections of the observations of the dependent variable onto the p-1 column vectors of the independent variables:

$$\hat{\mathbf{z}}_k = \mathbf{Z}^*(\mathbf{Z}^{*\prime}\mathbf{Z}^*)^{-1}\mathbf{Z}^{*\prime}\mathbf{z}_k$$

where $\mathbf{z}_k$ is the vector of observations of the mean-centered dependent variable and $\mathbf{Z}^*$ is the matrix of independent variables.

Error Theory

LPCA shares the assumption of regression analysis that the error lies in the same dimension for every observation, but LPCA is less restrictive in not assuming a priori what dimension. In regression the error term is assumed to be in the dimension of the dependent variable. In LPCA there is no prior assumption about the dimension; all variables can contribute to it. It is in this sense that LPCA acknowledges the possibility that more than one variable is latent or random or contains measurement error. The estimated error for LPCA lies orthogonal to the p-1 dimensional space of greatest variance.

The matrix of error $\hat{\mathbf{E}}$ is the projection of the mean-centered observations $\mathbf{Z}$ onto the partitioned matrix $\mathbf{A_2}$ of eigenvectors containing zeroes for the first p-1 eigenvectors:

$$\hat{\mathbf{E}} = \mathbf{Z}\mathbf{A_2}''\mathbf{A_2} = \mathbf{Z}(\mathbf{I} - \mathbf{A_1}''\mathbf{A_1}).$$

Regression Analysis assumes that there is no error, latency, or randomness in the p-1 dimensions of the independent variables; all is in the dependent variable, a strong assumption. While possible for LPCA, as previously mentioned, it is not assumed.

The error estimates for Regression are the projections of the dependent variable $\mathbf{z_k}$ onto the orthogonal complement of the column space of the independent variables:

$$\hat{z}_k = (\mathbf{I} - \mathbf{Z}^*(\mathbf{Z}^{*'}\mathbf{Z}^*)^{-1}\mathbf{Z}^{*'})\mathbf{z_k}.$$

Although LPCA and regression analysis intersect where the dependent variable is the observed multivariate dimension of least variance, regression analysis is not a subset of LPCA.

Covariance Requirements

The covariance assumptions of LPCA and Regression Analysis also provide insights into the relative assumptions. Whereas, the covariance matrix of interest in LPCA is the one of all variables, the covariance matrix of interest in Regression Analysis is the one of the independent variables. Both methods of linear analysis assume that p-1 of the eigenvalues do not equal zero. For Regression, these are the p-1 dimensions of the independent variables.

LPCA assumes that the two smallest eigenvalues do not equal each other--called the sphericity assumption. If any two linear combinations equal each other in total variability then any one of their infinite combinations is an equally good fit and LPCA cannot discriminate (Figure 4).

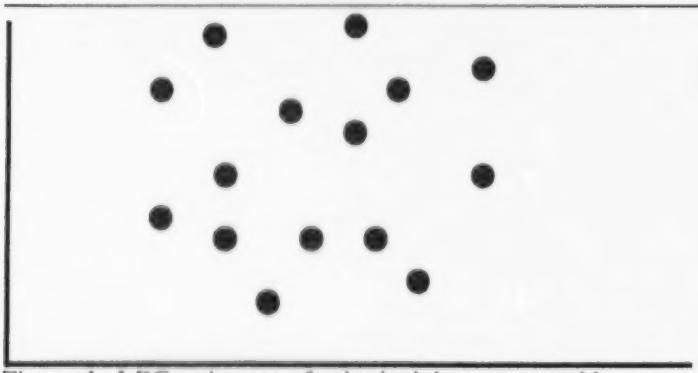

Figure 4. LPC estimates of spherical data are unstable

The sphericity test, described in the next section, checks this assumption.

On the other hand, Regression Analysis assumes that the ratio of the largest and smallest eigenvalues is not so large as to create problems of multicollinearity.

How Are LPCA Coefficients and Standard Errors Estimated?

To estimate a linear equation with LPCA:

(1) Compute eigenvectors and eigenvalues using the covariance matrix rather than the correlation matrix;
(2) Take the elements of the eigenvector associated with the smallest eigenvalue as the estimates of the linear coefficients of the respective variables; and
(3) Compute the intercept by summing the products of each of the LPC coefficients and means of the variables.

Where n is the number of observations, it does not matter whether you use 1/n or 1/(n-1) in computing the covariance matrix, the eigenvectors remain the same (Flury, 1988, p. 14 and 17).

Anderson (1963) developed the asymptotic theory of statistical inference for Common Principal Components .[2] His result, adapted to LPCA, provides an estimate of the standard error:

$$s(a_{mp}) = \left[\frac{1}{n} l_p \sum_{i=1}^{p-1} \frac{l_i}{(l_i - l_p)^2} a_{mi}^2\right]^{\frac{1}{2}}, \text{ for m = 1 to p, where}$$

n is the number of observations, l_i is the ith eigenvalue, a_{mi} is the mth element of the eigenvector associated with the ith eigenvalue, and p is the index of the smallest eigenvalue. The multivariate nature of the estimates makes a Bonferroni adjustment appropriate. The sphericity test is implicitly included.

Flury (1988, pp. 29-30) provides an explicit test of sphericity--the null hypothesis that the smallest eigenvalue is not smaller than the next smallest one:

$$S(l_{p-1}, l_p) = \frac{2n \log (l_{p-1} + l_p)}{2(l_{p-1})^{1/2} l_p} \approx X_2^2.$$

The tests of statistical significance and confidence intervals for the LPC estimates and sphericity of the smallest eigenvalue do not depend on the normality of the data. They apply as long as the fourth moments of the data exist (Muirhead, 1982, p. 19, Theorem 1.2.17). Also, Diaconis and Efron (1983), Stauffer et al. (1985), and Daudin et al. (1988) have applied non-parametric bootstrap methods to infer statistical significance and confidence intervals for non-normal principal components (Flury, 1988).

An Example of LPCA Estimation of Academic Activity and University Faculty

Faculty members divide their academic time between instructing students of differing types and undertaking scholarship leading to publications. The measurements of the numbers of students of differing types, faculty, and publications are subject to variability in occurrence and measurement.

We want the best total linear fit so we will use LPCA. Common Principle Components Analysis provides us with the eigenvalues and associated eigenvectors displayed in Table 2.

---

[1]Few, if any, major packages have the inferential statistics for Principal Components Analysis, which could also have been used for LPCA. Anderson's asymptotic statistics are reasonably straightforward to compute from the eigenvalues and principal components which most packages do provide.

Table 2
Estimates of Eigenvalues and Eigenvectors of the Covariance Matrix of Data Relating Faculty Members and Academic Activity

	One	Two	Three	Four
Eigenvalues	49,331,217	2,965,478	132,270	16,667
Variance Proportion	0.941	0.057	0.003	0.000
Principal Components				
Faculty	-0.054	-0.056	0.159	-0.984
Graduate Students	-0.196	-0.949	-0.245	0.026
Undergraduates	-0.978	0.204	-0.003	0.041
Articles	-0.045	-0.233	0.956	0.17

The LPC estimates are the eigenvectors associated with Eigenvalue Four. One minus the ratio of the smallest eigenvalue to the sum of all of the eigenvalues provides an estimate of the proportion of covariance explained. Our model explains effectively 100 percent of the total variance.

Table 3 displays the normalized LPC estimates and their Regression equivalents.

Table 3.
LPCA and Regression Coefficients, Standard Errors, T-ratios, and Proportions of Variance Explained for Data Relating Faculty Members and Academic Activity

	Coefficient (Normalized)	Coefficient (Original)	s.e.	T-ratio (Original)	Proportion Explained
LPCA					1.000
Faculty	1.000	-0.984	0.004	-245.134	
Graduate Students	0.026	0.026	0.008	3.419	
Undergraduates	0.042	0.041	0.001	27.901	
Articles	0.173	0.170	0.024	7.052	
Regression					
Faculty	1.000	1.000			
Graduate Students	0.031	0.031	0.007	4.340	
Undergraduates	0.042	0.042	0.002	28.000	
Articles	0.152	0.152	0.022	6.890	

The estimates normalized per unit of faculty reflect the proportion of faculty time needed for an additional academic activity of each kind. An additional graduate student requires an additional 2.6 percent of faculty time. An additional undergraduate requires an additional 4.2 percent. And, an additional article requires an additional 17.3 percent. The amount of time per article is probably proportional but overstated because the article counts are from citation indexes that include only the most frequently referenced journals. There is little distinction between the Regression and the LPC estimates in this model, which would not be the case if the dimension of faculty did not contain so little unexplained error. It provides a good example of how similar the two methods can be.

Summary and Conclusion

Least Principal Components (LPC) are robust, consistent, and sufficient maximum likelihood estimates of the best total linear fit to observed data. They are more appropriate than Regression estimates when (1) the smallest eigenvalue exists and is distinct from the next smallest and (2a) the variability to minimize is in more than one variable, or (2b) multicollinearity is a problem. They are as easy to compute as Common Principal Components, because they are principal components. Anderson (1963) provides a theory of inferential statistics for Principal Components that can be used in computing significance levels and confidence intervals for LPC as well. Boot-strap approaches have also been developed.

References

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. Annals of Mathematical Statistics, 34, 122-148.

Daudin, J. J., Duby, C., & Trecourt, P. (1988). Stability of principal component analysis studied by the bootstrap method. Statistics, 19, 241-258.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 248(May), 116-130.

Draper, N. R., & Smith, H. (1981). Applied Regression Analysis (2d ed.). New York: John Wiley & Sons.

Flury, B. (1988). Common Principal Components and Related Multivariate Models. New York: John Wiley & Sons.

Fuller, G. (1967) Analytical Geometry (3d ed.). Reading, MA: Addison-Wesley.

Greenberg, E. (1975). Minimum variance properties of principal component regression. Journal of the American Statistical Association, March (70), 194-197.

Kendall, M, Stuart, A., & Ord, J. K. (1977). The Advanced Theory of Multivariate Statistics (4th ed.), Vol. 3. New York: Macmillan Publishing.

Kendall, M. G. (1961). A theorem in trend analysis. Biometrika, 48, 224.

Muirhead, R. (1982). Aspects of Multivariate Statistical Theory. New York: John Wiley & Sons.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine, series 6, 2, 559-572.

Stauffer, D. F., Garton, E. O., & Steinhorst, R. K. (1985). A comparison of principal components from real and random data. Ecology, 66, 1693-1698.